Université Paris Diderot - Paris 7 2017-2018

Master 2 Biologie-Informatique/ Bioinformatique

Etude de la fonction et des mécanismes d'évolution des séquences répétées centromériques chez les Primates

Sarah Kaddah

Tuteur : Loïc Ponger

Structure et Instabilité des Génomes

MNHN - CNRS UMR 7196 / INSERM U1154 - Sorbonne Universités

Muséum national d'Histoire naturelle, 43 rue Cuvier 75005 PARIS

Remerciements

Merci à Namrod pour toute la partie sur la bibliographie. Retrouvez ses questions FAQ qui ont permis la rédaction de cette partie.
Merci à f-leb, LittleWhite et Metalman pour leurs conseils et la relecture. Merci à ced et jacques_jean pour la correction orthographique et typographique.

Table des matières

1 Introduction

1.1 Le centromère

Bibliographie :

1-> kinetochore

2-> CENP-A

3-> l'ADN satellite

Le centromère est une structure chromatinienne caractérisé par la présence de CENP-A. Cette protéine, très conservée au cours de l'évolution, est un variant de l'histone H3. Son rôle est de fixer la position du kinétochore par un mécanisme encore peu connu. En effet, le centromère est le site d'assemblage du kinétochore, un ensemble d'ADN et de protéines. Il permet l'attachement du fuseau mitotique pour la ségrégation des chromosomes durant la division cellulaire chez les eucaryotes. Le centromère et les protéines impliquées sont relativement bien conservés. Au contraire, l'ADN sous-jacent est très diversifié et l'organisation varie d'un taxon à l'autre. Cependant, une caractéristique commune est retrouvée chez toute les espèces : de l'ADN centromérique répété en tandem nommé ADN satellite. Ces répétitions sont issues d'événements d'amplification, tels les crossovers inégaux, la conversion de gènes, les cercles roulants ou la transposition de séquences.[Malik and Henikoff, 2002 ; Plohl et al. 2012] Ces séquences représentent 5% du génome. Les répétitions s'étendent de 7pb à 3,2kb avec des séquences de 145-180kb le plus souvent.

1.2 L'ADN α-satellites

-> première mise en évidence des AS

->théorie gradient de l'âge

-> travaux sur le gorille à dev

L'ADN satellite chez les Primates est connu sous le nom d'α-satellite. Ces séquences centromériques répétées en tandem sont riches en AT. -> article sur la première découverte

Des études chez l'Homme propose un modèle évolutif. La répartition des α-satellites suivrait une répartition spécifique selon l'âge des familles. Les familles les plus jeunes s'insèrent au cœur du centromère, repoussant les familles les plus anciennes jusqu'aux regions voisines,

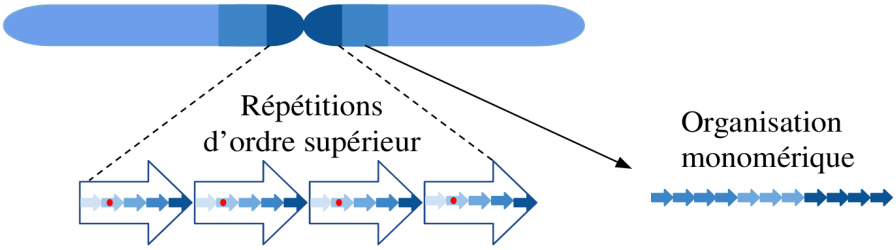

FIGURE 1 – **Organisation spatiale des α-satellites :** Le coeur du centromère (bleu foncé) est organisé en répétition d'ordre supérieur. Le péricentromère(bleu clair) a une organisation monomérique. Un monomère d'une même famille est représenté par une petite flèche de même couleur. Les points rouges représentent les sites de fixation à CENP-B ou pJα.

appelé péri-centromère.

->Article Shepelev

->Est-ce que je peux utiliser du conditionnel ? OU est-ce que cette théorie est confirmée ?

Un monomère a une longueur de 171pb et il peut être répété des milliers de fois. Les monomères peuvent être répartis en famille selon leur similarité, les séquences ayant un taux d'identité supérieur à 70%. Ces séquences ont soit une organisation monomérique soit une organisation en répétition d'ordre supérieur (Fig. 1). Dans le premier cas, les séquences d'une même famille sont répétés en tandem. Dans le deuxème cas, une suite de monomères appartenant à différentes familles forme une unité, qui elle est répétée en tandem.

Ces séquences peuvent avoir un site de liaison à la protéine centromérique CENP-B un motif spécifique de 17pb. Cette protéine, qui reconnaît et se fixe sur l'ADN, serait présente chez de nombreuses familles de Primates. La protéine pJα, une protéine peu caractérisée, reconnaît un motif qui remplace celui de CENP-B.

Les α-satellites ont essentiellement été étudiées chez l'homme. Modèle évolutif avec les centromères en expansion. Une hypothèse concernant l'âge des séquences découle de ces recherches : les séquences les plus récentes apparaissent au coeur du centromères, déplançant les plus anciennes au péricentromère. D'autres études chez le gorille ont été faites. Le rôle des α-satellites est encore mal connu.

1.3 Le sujet de stage

Peu d'informations sur les α-satellites figurent chez les autres espèces de primates, et aucune relation inter-espèce n'a été réalisée. Les études se basant sur un séquençage haut-débit est appliqué chez l'homme (cité ci-dessus) et chez le Gorille. [compléter] L'équipe d'accueil

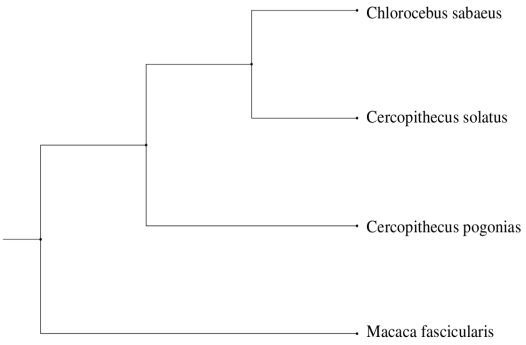

FIGURE 2 – **Arbre phylogénétique des espèces analysées.**[1]

de mon stage "ADN répété, Chromatine, Evolution" ou ARChE, a récemment développé une approche de séquençage haut débit, ciblée sur les séquences α-satellites chez deux espèces : *Cercopithecus solatus* et *Cercopithecus pogonias*.

Les méthodes basées sur l'alignement et la phylogénie sont très limitées pour étudier ces séquences. Elles ne permettent pas de traiter des jeux de données conséquents, or un monomère peut avoir des milliers de copies dans un seul génome. De plus, ces méthodes non-objectives ne permettent pas de faire des comparaisons entre espèces.

Pour remédier à ce problème, une méthode de classification automatisée des α-satellites a été implémentée en R en 2016, puis améliorée en 2017 en Python dans le laboratoire. Ce programme permet de traiter des centaines de milliers de séquences, quelque soit le nombre ou la taille des familles. De plus, cette méthode est objective et peut être appliquée à plusieurs espèces, permettant ainsi une comparaison inter-espèce des familles α-satellites.

L'objectif de ce stage est de comprendre la fonction des α-satellites, notamment en caractérisant les familles issues de cette classification chez quatre espèces proches de primates. Parmi ces espèces, les deux Cercopithèques séquencés dans le laboratoire permettront d'avoir un avis objectif sur la méthode de classification. Dans une deuxième temps, les mécanismes d'évolution pourront être déduits à partir d'une comparaison inter-espèce révélant les différences et les familles communes.

2 Matériel et méthode

2.1 Choix des espèces

Les critères de sélection dépendent de la disponibilité des séquences de qualité. Deux espèces du laboratoire sont choisies, le *C. solatus* et *C. pogonias*, et deux espèces proches (Fig. 2),

le *Macaca fascicularis* et le *Chlorocebus sabaeus*.

2.2 Méthode de classification

2.2.1 Principe

Cette méthode [2] répartit des séquences α-satellites en familles selon la similarité. La classification est hiérarchique dichotomique. Au départ, une table contenant les fréquences des 5-mers est calculée pour chaque monomère. Ensuite une boucle itérative est exécutée pour séparer les séquences en groupes tant que les nouveaux groupes formés sont divisibles.

2.2.2 Répartition itérative

Une Analyse en Composante Principale (ACP) est effectuée sur la table des fréquences des 5-mers afin de réduire les dimensions du jeu de données et d'obtenir des variables indépendantes. Des distances euclidiennes sont calculées entre toutes les paires de séquence dans l'espace défini par les premières composantes de l'ACP.

A partir du calcul de distance, les séquences sont séparées en deux classes en utilisant la classification hiérarchique basée sur la méthode de Ward. Cette méthode maximise l'inertie interclasse. La classification hiérarchique fait un usage important de la mémoire. Par conséquent, pour traiter des jeux de données importants de plus de 100 000 séquences, l'Analyse Discriminante Linéaire , une méthode d'apprentissage, est utilisée sur un sous jeu de données formé par de 100 000 séquences tirées aléatoirement, dans ces analyses. Le modèle construit est alors appliqué sur toutes les séquences.

2.2.3 Double-validation d'un sous-groupe

Le premier critère de validation est la taille du sous-groupe. Si un groupe atteint 100 séquences, il n'est pas redivisé. Le deuxième critère de validation s'appuie sur le *matepair*. Ce terme correspond à la proportion de monomères ayant son plus proche voisin dans la même classe, se basant sur les distances euclidiennes calculées auparavant. Des valeurs *matepairs* élevées (proches de 1) indiquent des sous-groupes bien homogènes et séparés validant la classification tandis qu'un seuil *matepair* plus faible (proche de 0) entraîne plus de classes.

Un seuil de *matepair* est fixé à 0.90, pour avoir des groupes homogènes. Si au moins une des valeurs de *matepair* est au-dessous de ce seuil, les sous-groupes sont considérés comme

formant un seul groupe et le groupe initial est sauvegardé comme une famille unique. Si les *matepairs* sont au-dessus d'un certain seuil, les deux sous-groupes sont ajoutés séparément à la file pour être potentiellement redivisés ultérieurement.

2.3 Analyse des séquences

2.3.1 Alignements

L'alignement des séquence est fait avec muscle [3] et SeaView [**?**], un éditeur d'alignements multiples.

2.3.2 Phylogénie

La phylogénie est construite avec la méthode du maximum de vraisemblance (PhyML) [**?**]. Le modèle F84 est utilisé pour la construction de l'arbre. Le support de branche est aLRT (SH-like). La fréquence d'équilibre nucléotidique est optimisée. Le ratio de transition et de transversion est fixé à 4. Aucun site est considéré comme invariable. Le taux de variation à travers le site est optimisé. Les opérations de recherche d'arbre est NNI et l'arbre de départ est défini avec la méthode de Neighbor-Joining [**?**] avec une topologie optimisée.

2.3.3 Consensus

Les consensus sont obtenus avec des scripts développés par l'équipe. Les motifs CENP-B (TTCGTTGGAA[AG]CGGGA), PJα (TTCCTTTT[CT]CACC[AG]TAG) et pKβ (CTATAGGGCCAAAG-GAA) ont été identifiés avec le logiciel fuzznuc (package EMBOSS) [4] et en autorisant 2 différences au maximum par rapport au consensus.

3 Résultat

3.1 Caractérisation intraspécifique des familles

3.1.1 Identification des familles

A l'issue de la classification, seules les familles ayant plus de 100 séquences sont conservées pour l'analyse. Elles sont qualifiées de grandes familles. Le nombre de familles conservées diminue considérablement après élimination des petites familles (<100 séquences) mais des

Espèce	*C. solatus*	*C. pogonias*	*C. sabaeus*	*M. fascicularis*
Nb. seq. au total	105 529	112 902	29 842	235 535
Nb. fam.	564	132	338	3642
Nb. grandes fam.	12	13	43	105
% seq. ignorées	3.97	1.29	10.89	10.75

TABLE 1 – Résumé du jeu de données et des résultats préliminaires de la classification.

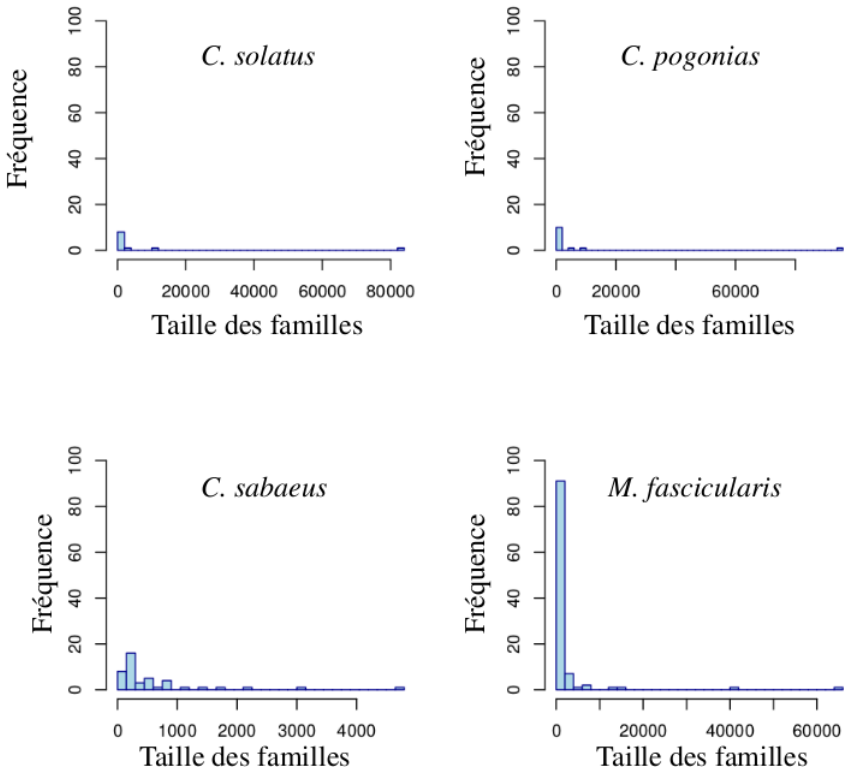

FIGURE 3 – **Distribution des familles en fonction de la taille.**

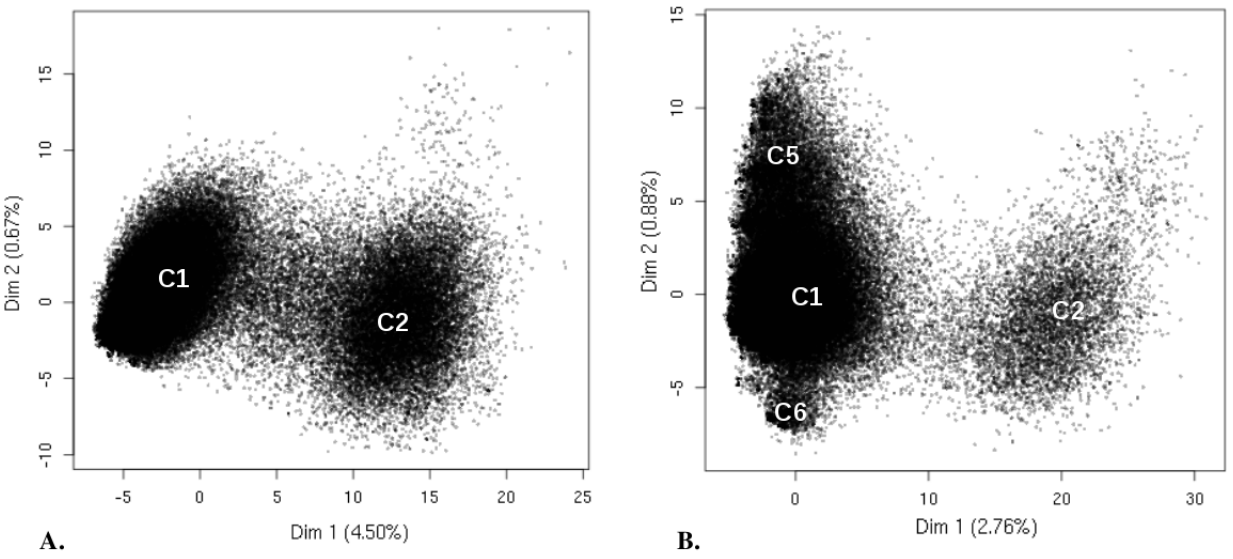

FIGURE 4 – **Caractérisation visuelle des familles α-satellite chez *C. solatus* et *C. pogonias* à partir d'ACP :**Le nom des familles est indiqué sur les graphiques. Un point représente un monomère. **A.** *C. solatus*. **B.** *C. pogonias*.

familles ne représentent que 11% du jeu de données au plus (Tableau 1). Les *C. solatus* et *C. pogonias* ont une dizaine de familles, *C. sabaeus* en a 40, et le *M. fascicularis* en a 105. La distribution des familles est similaire chez les quatre espèces. Les familles les plus petites sont très nombreuses et les familles les plus grandes sont peu fréquentes (Figure 3). *C. solatus* possède une grande famille de 80 000 séquences, *C. pogonias* de 94 000 séquences, *M. fascicularis* possède quatre familles de plus de 10 000 séquences. La plus grande famille de *C. sabaeus* fait 5000 séquences. Ces familles se démarquent des autres familles.

Les espèce *C. solatus* et *C. pogonias* sont analysées dans un premier temps pour comparer la classification automatisée avec la classification expérimentale du laboratoire.Expérimentalement, 6 familles α-satellites ont été déterminées chez les Cercopithèques. Ces deux espèces partagent deux grandes familles monomériques, C1 et C2, de l'ordre de plusieurs milliers de séquences, et deux familles formant un dimère, C3-C4, de l'ordre d'une centaine de séquences chacune. *C. pogonias* possède les familles supplémentaires C5 et C6. Ces familles ont été définies à partir d'une méthode visuelle établie d'après une ACP (Fig. 4).

Bien que le nombre de grandes familles soit relativement proche entre ces deux espèces, les résultats diffèrent significativement (Tableau 2). Toutes les familles chez *C. solatus* sont retrouvées : 11 familles forment la famille C2, une famille forme la famille C1 et les familles C3 et C4 sont retrouvées dans des petites familles d'environ 80 séquences chacune. Chez *C. pogonias*,

Espèces / Fam. exp.	*C.solatus*	*C.pogonias*
C1	1	10
C2	11	1
C3	(1)	1
C4	(1)	(1)
C5	-	1
C6	-	0

TABLE 2 – **Résumé du tableau de contingence :** Comptage des familles issues de la classification et leur répartition théorique dans les familles expérimentales (C1 à C6). Les valeurs entre parenthèse sont des petites familles (< 100 séquences) qui ne sont pas prises en compte dans le reste des analyses.

toutes les familles sont retrouvées sauf la famille C6. La famille C1 est répartie en 10 familles, les familles C2, C3 et C5 sont retrouvées entièrement, la famille C4 est également retrouvée sous la forme d'une petite famille de 86 séquences. Chez le *C. solatus*, la famille C2 est divisée en plusieurs familles et la famille C1 est retrouvée dans une seule famille. La situation inverse est retrouvée chez *C. pogonias*.

Pour visualiser cette comparaison, des couleurs sont assignées aux familles issues de la classification automatisée. Ces couleurs sont superposées aux résultats expérimentaux en noir. Chez *C. solatus*, la famille C1 est entièrement retrouvée dans une famille. La famille C2 est répartie en plusieurs familles. Deux familles intermédiaires (rouge et bleue) sont visibles entre la famille C1 (vert) et C2 (orange). Elles ne sont pas distinctes. Une famille supplémentaire (turquoise) se démarque. Pour confirmer cette division de la famille C2, la visualisation de l'ACP des 5-mers est observée en fonction des composantes 1 et 3. Les familles intermédiaires sont toujours confondues, contrairement à la famille turquoise qui forme une famille à part entière. Pour certifier ce fait, l'arbre construit atteste que chaque famille est bien retrouvée, notamment les familles intermédiaires qui forment bien deux familles. Chez *C. pogonias*, les familles C2, C4 et C5 sont bien retrouvées. La famille C6 se fond dans la famille C1 (en vert). La famille C1 est divisée en deux familles supplémentaires (rose et violet). La visualisation des composantes 1 et 3 de l'ACP ne permet pas de trancher sur la classification. L'arbre montre que les familles en rose et violet sont très proches.

3.1.2 Motifs fonctionnels

La protéine CENP-B est présente chez toutes les espèces, mais les Cercopithèques ne possèdent pas son site de liaison. Cette assertion est vérifiée par l'absence de pourcentage pour

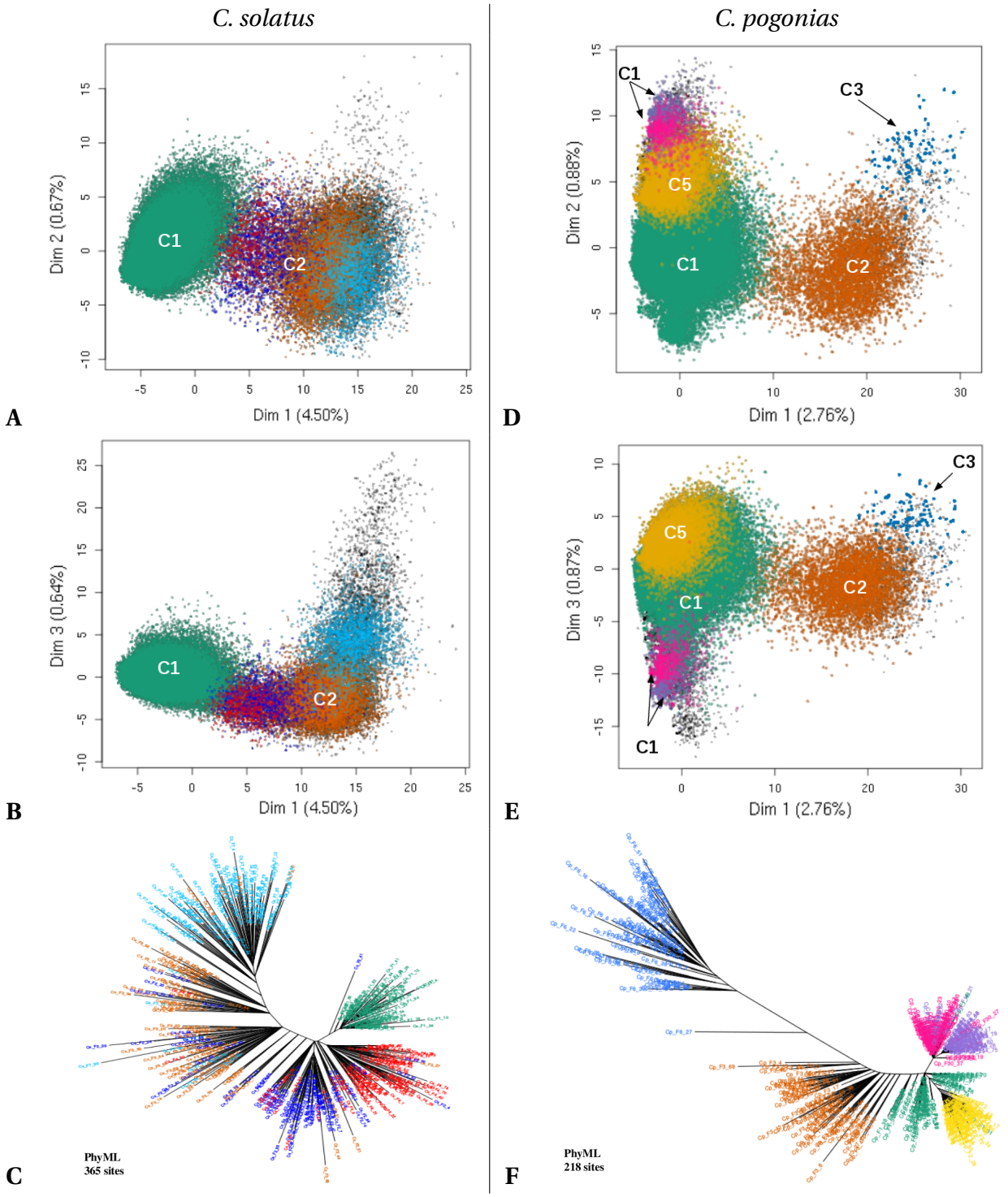

FIGURE 5 – **Représentation des plus grandes familles issues de la classification automatisée :** Les plus grandes familles issues de la classification automatisée sont indiquées sur les représentations de l'ACP des 5-mers. **A.** Composantes 1 et 2 de l'ACP. C1 est en vert, C2 est en orange, rouge, bleu et turquoise. **B.** Composantes 1 et 3 de l'ACP. **D.** Composantes 1 et 2 de l'ACP. De même pour *C. pogonias*. C1 est en vert, violet et rose, C2 est en orange, C4 est en bleu clair, C5 est en jaune. **E.** Composantes 1 et 3 de l'ACP. **C. et F.** Représentation des familles issues de la classification sous forme d'arbres. Les couleurs des familles sont respectivement conservées et 100 séquences alétoires par familles sont sélectionnées pour la construction de l'arbre.

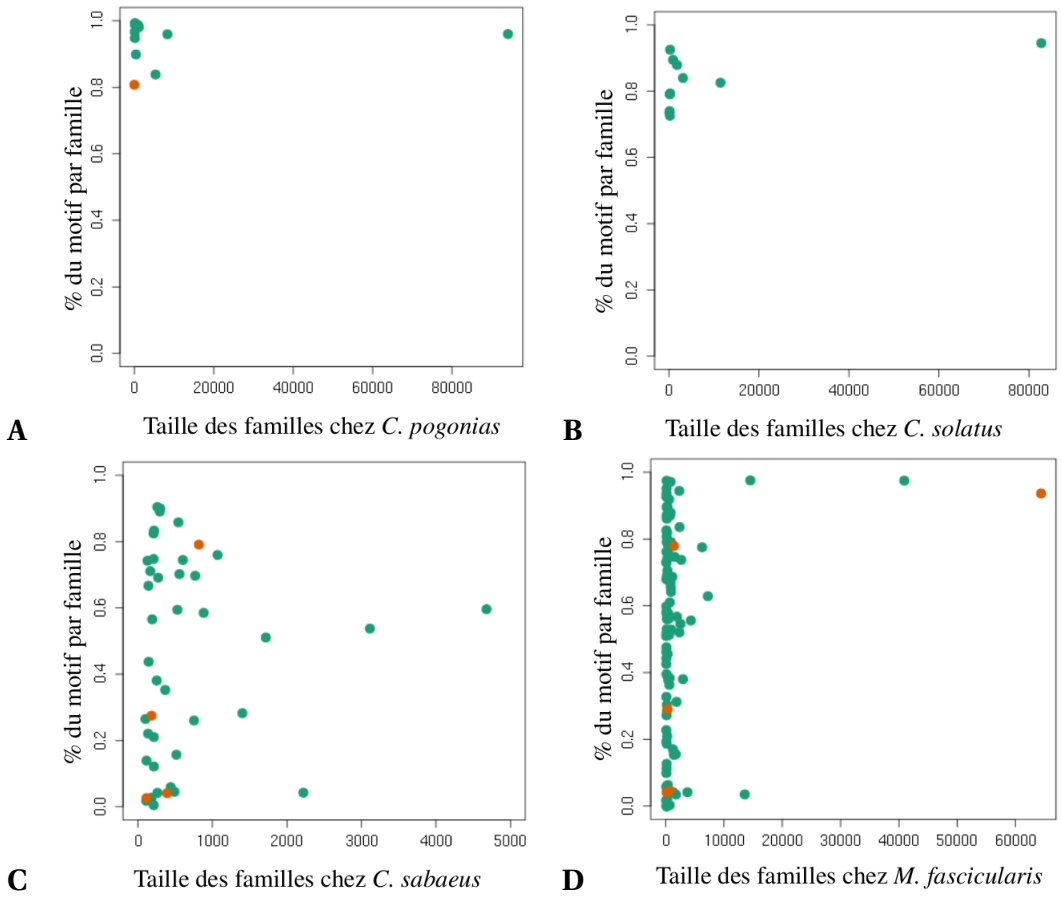

FIGURE 6 – **Présence des motifs CENP-B, pJα ou pKβ par famille :** Le pourcentage de séquences par famille ayant le motif pJα est en vert, pKβ en orange et CENP-B en bleu. Chaque famille est représentée en fonction de sa taille.

ce motif dans les graphes. Le *C. sabaeus* et le *M. fascicularis* n'ont pas ce motif non plus. Le *M. fascicularis* a 7 familles sans motifs.

Au contraire, la protéine pJα est très présente chez *C. solatus* (à 90% en moyenne) et *C. pogonias* (à 85% en moyenne). La plus grande famille C1, ayant plus de 80 000 séquences chez ces deux espèces, se démarque avec un pourcentage à 95%.

Le *M. fascicularis* ont des pourcentages pour le motif pJα qui varie entre 1% et 97%. Cette espèce a 20 familles avec le motifs à plus de 80% dont une famille de 40 000 séquences et une autre de 14 000 séquences, 35 famille ayant un pourcentage entre 50% et 80% exclu, et 40 famille ayant le motif à moins de 50%.

Le motif pKβ est présent lorsque pJα est absent de la famille. Il est absent chez *C. pogonias*. Seule la famille C3 a ce motif à 80% chez *C. solatus*. Le *C. sabaeus* a deux familles avec le motif, l'une à 79% et l'autre à 29%. Le *M. fascicularis* a trois familles avec ce motif. La famille ayant le motif à 93% est une grande famille de 64 000 séquences. Les deux autres familles ont le motif à 0.77% et 0.29%.

3.1.3 Similarité entre familles

C. solatus et *C. pogonias* ont un pourcentage de divergence qui est relativement faible, ne dépassant pas les 15%. La famille C1 a un pourcentage de divergence très faible de 5%. Cette grande famille de plus de 80 000 séquences est donc probablement une famille récente. *C. pogonias* a d'autres familles qui ont un pourcentage en dessous de 5%. *C. sabaeus* a les pourcentages de divergence les plus élevés, allant jusque 28%, quelque soit la taille de la famille. Cette espèce a des familles anciennes d'α-satellites ou alors cette divergence est le résultat d'erreurs de séquençages. Le *M. fascicularis* a des poucentages de divergence qui varient entre 5% et 27%. 3 grandes familles ayant plus de 10 000 séquences ont un pourcentage de divergence qui se rapproche de 5%. Ces familles sont probablement récentes.

Les tailles moyennes des consensus sont de 184 nucléotides pour *C. pogonias*, 206 pour *C. solatus*, 203 pour le *C. sabaeus* et X pour le *M. fascicularis*

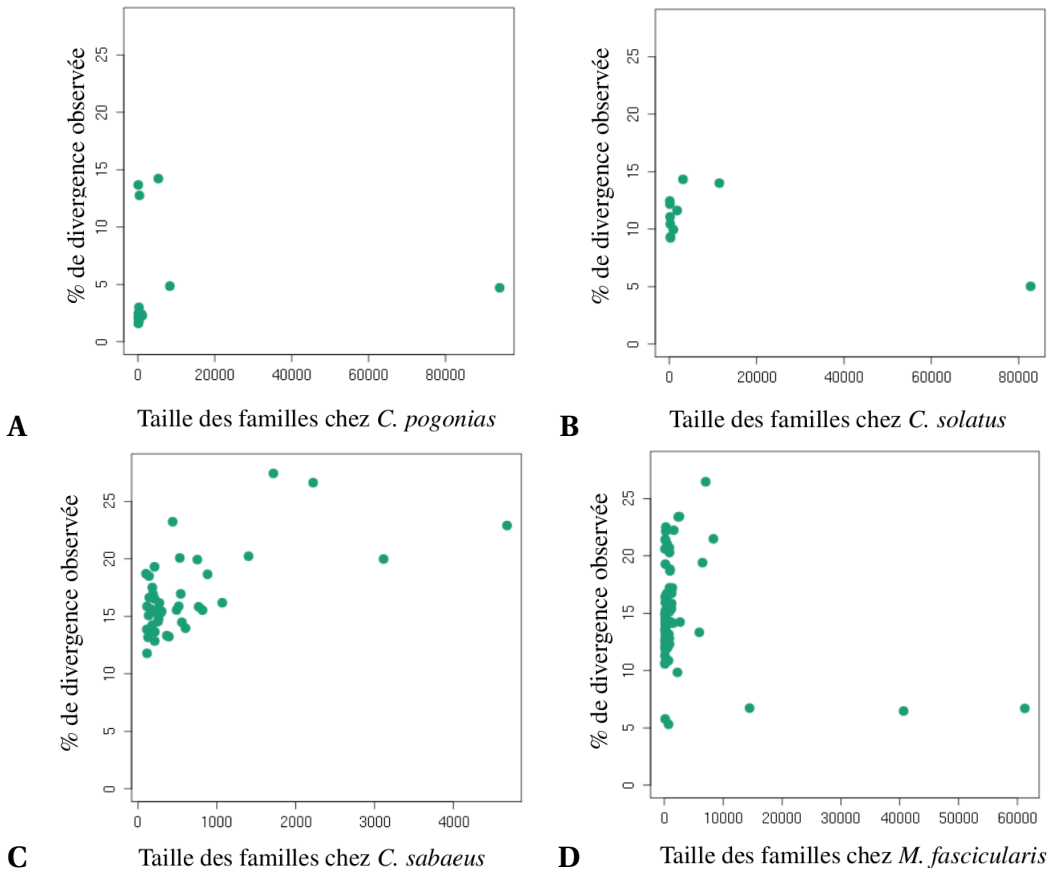

FIGURE 7 – **Pourcentage de divergence observée au sein d'une famille :** Un point correspond à une famille et représente le pourcentage de divergence en fonction de la taille de la famille.

3.2 Comparaison inter-espèce

3.2.1 Répartition des super-familles

Pour étudier les mécanismes d'évolution des α-satellites, une classification inter-espèce ou "super-classification" permet de comprendre les différences entre espèces. Pour chaque espèce, 100 séquences par grande famille α-satellite sont tirées aléatoirement. La taille minimale d'une super-famille est fixée à 20 séquences. A l'issue de cette super-classification, 139 familles sont obtenues au total, dont 92 grandes familles. Les petites familles de moins de 20 séquences, soit 1.33% de ce jeu de données, ne sont pas prises en compte dans l'analyse.

Les différentes espèces partagent des super-familles, mais certaines sont spécifiques. Le *M. fascicularis* possède 15 familles spécifiques. Le *C. pogonias* possède 8 super-familles qui lui sont spécifiques. L'une de ces super-familles appartient a un sous-groupe de C6 de 200 séquences. Les autres super-familles sont des sous-familles de la famille C1. Le *C. solatus* possède trois super-familles qui lui sont spécifiques. Or ces familles sont des sous-familles de la famille C2. Les familles C1 et C2, sont des familles qui sont initialement décrites communes aux *C. pogonias* et C. solatus. Malgré des différences, la super-famille 23 qui est partagée entre ces deux espèces appartient également à la famille C2.

Cependant le *C. sabaeus* ne possède pas de famille spécifique. Il partage 35 de ses super-familles avec le *M. fascicularis*. La super-famille 24 est commune à *C. solatus*, sabaeus et au macaque. Dans cette super-famille, les 118 séquences appartenant à *C. solatus* font parti de la famille C2. Cette famille est donc une famille ancestrale partagée par ces 4 espèces.

La famille C2 fait l'objet d'une séparation intéressante. Une partie est commune à *C. pogonias* et *C. solatus* tandis qu'une autre est commune à *C. pogonias*, *C. sabaeus* et *M. fascicularis*. Pour vérifier les résultats de cette classification, 30 séquences de chacune des familles C2 de *C. solatus* et *C. pogonias* sont tirées aléatoirement et un arbre est construit pour voir si cette division est retrouvée , puis un autre arbre avec les familles des deux autres espèces supposés de la famille C2 est également construit(résultat non montré). Effectivement, toutes les regroupements de famille sont retrouvés.

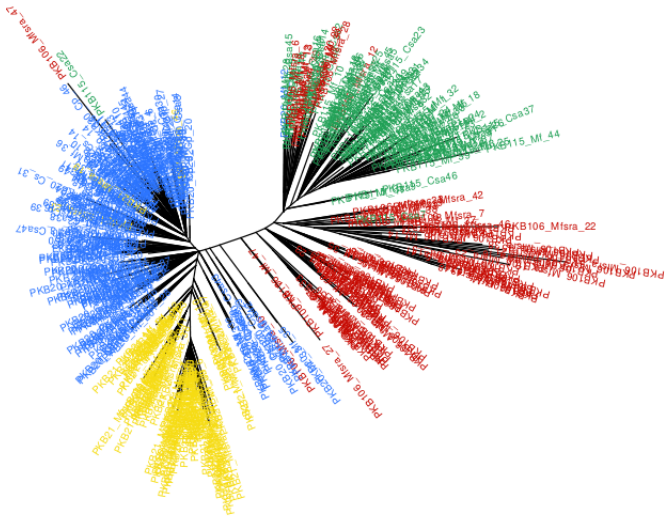

FIGURE 8 – **Abre des super-familles pKβ :** La super-famille 106 est en rouge, 115 en vert, 20 en bleu et 21 en jaunes.

3.2.2 Origine du motif pKβ

Pour poursuivre l'étude sur les familles communes entre espèces, le regroupement des familles α-satellites ayant le motif pKβ ou "familles pKβ" attire l'attention. D'abord sont analysées les familles qui ont le motif à plus de 75%. Elles représentent au mieux les familles pKβ. Les familles qui expriment moins le motif sont analysées par la suite, pour voir si malgré la faible présence du motif, elles se rassemblent en une ou plusieurs super-familles. Le premier groupe de familles pKβ est constitué de la famille C3 de *C. pogonias*, la famille 3 de *C. sabeus*, et les familles 1 et 17 de *M. fascicularis* qui expriment le motif à respectivement 80%, 79%, 94% et 80%. Ces familles s'assemblent en deux super-familles. La super-famille 21 est spécifique au macaque, regroupant la famille 1. La super-famille 20 regroupe les familles C3 de *C. pogonias*, la famille 3 de *C. sabeus* et la famille 17 de *M. fascicularis*. Toutes ces espèces ont donc une "famille C3" en commun, peut-être sous forme de dimère C2-C3. Le deuxième groupe de famille pKβ est formé de la famille 27 de *C.sabeus* et les familles 47 et 294 du *M. fascicularis* qui ont le motif à 27%, 16% et 41% respectivement. La super-famille 106, regroupant la famille 294, est spécifique de *M. fascicularis*. Les deux familles restantes, qui ont un faible pourcentage du motif, se regroupent pour former la super-famille 115. De plus, ces super-familles sont uniquement consitutées de familles pKβ et ne se mélangent pas aux autres familles. Un arbre composé de toutes les super-familles pKβ est construit pour voir comment elles s'assemblent. Le jeu de données est construit à partir de 50 séquences par famille pKβ par espèce tirées alétoirement (Figure 8) et contiennent la famille C3 de *C. solatus*, dont la famille était trop petite pour être

prise en compte. Comme les autres "familles C3", elle se range dans la super-famille 20.

4 Discussion

5 Conclusion

Références

[1] Cacheux. Evolutionary history of alpha satellite dna in cercopithecini.

[2] Florence Jornod. Master's thesis, Paris Diderot, 2016-2017.

[3] Robert C. Edgar. Muscle : multiple sequence alignment with high accuracy and high throughput, 2004.

[4] Peter Rice, Ian Longden, and Alan Bleasby. Emboss : the european molecular biology open software suite, 2000.

Résumé

Votre résumé commence ici... ...

Abstract

Abstract begins here... ...